This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics

Driver Identification Based on Wavelet Transform Using Driving Patterns

Byung Il Kwak, *Student Member, IEEE,* Mee Lan Han, *Student Member, IEEE,* and Huy Kang Kim

***Abstract*—The modern automotive system, based on the convergence of information and communication technologies, is equipped with various functions to ensure vehicle safety and convenience of the driver. A driver-identification technology is an effective method to perform vehicle-theft detection. It can also provide customized driver-personalization services, such as healthcare or insurance. In the present study, we propose and evaluate a driver-identification method based on wavelet transform by performing driving-pattern analysis for each driver. We compare the performances of three different machine-learning algorithms, namely Support Vector Machine, Random Forest, and XGBoost for performing driver identification. The proposed method is applicable to both binary and multi-class classifications for the driving data of five drivers. In the case of motorway, the XGBoost classifier identifies each driver and delivers an accuracy of up to 96.18% in binary classification and an accuracy of 91.6% in multi-class classification. Moreover, in the case of an urban road, the SVM classifier achieves an accuracy of up to 95.07% in binary classification and accuracy of 89.06% in multi-class classification. The proposed method provides a context for a better understanding of the association between driver behavior, which is an in-vehicle event, and mechanical reactions. Our results shall help researchers to broaden the understanding of driver identification using in-vehicle data.**

***Index Terms*—Anti-theft, driver identification, driving pattern analysis, machine learning, wavelet transform.**

I. Introduction

AUTOMOTIVE technology is experiencing rapid advances in terms of providing both convenience and safety to drivers. Novel advanced automotive technologies are based on a computing platform connected to an external network. However, these state-of-the-art technologies increase the exposure to potential internal/external threats in the computing platform environment and introduce vulnerabilities to the in-vehicle cutting-edge technology. Moreover, novel technologies enhance the risk of taking control of the vehicle without permission from the vehicle owner. These aspects considerably deteriorate driver safety and vehicle security. Vehicle theft incidents have occurred frequently in recent years as examples of breaches in vehicle security, and most of these incidents were due to malicious use of computer technology [1]. Furthermore, as an important example of the attack vector, in a 2015 study, Charlie Miller and Chris Valasek succeeded in remoted control of in-vehicle functions by hacking Bluetooth, telematics, and the MP3 parser of the radio [2].

In-vehicle vulnerabilities reported to date include stealing, for the purpose of achieving financial benefit, a device embedded in the vehicle or the vehicle itself. Moreover, when a connected vehicle is in an autonomous driving environment, existing vulnerabilities might magnify the threat to driver safety and vehicle security. Currently, solutions such as smart keys, smart alarms, and GPS tracking are implemented to prevent vehicle theft or detect problems to protect against it. To ensure vehicle security in terms of theft, extensive research has been conducted on monitoring the driving status by extracting the biometric information of the driver. In addition, a detection method using a mechanical device has been proposed. In the present study, we introduce a method of identifying the driver by focusing solely on his/her driving pattern, without considering the driver's biometric information. The proposed method is based on the implicit authentication through the driver's driving style. Drivers exhibit different driving patterns depending on their age, gender, and personality. These driving patterns can serve as an appropriate indicator for driver authentication.

In the remainder of this section, we discuss relevant approaches and contributions of our method to driver identification using the existing classification algorithm. Section II provides a review of the literature related to our work. Our method of driver identification based on the wavelet transform is presented in Section III. The detailed implementation of our proposed approach and its evaluation are described in Section IV. Finally, Section V provides discussion of the results, conclusion, and directions of further research.

The main contributions of the present study are as follows:

1) We designed the driving dataset as a training dataset, while the testing dataset was based on previous trip data. Splitting of the dataset based on trip data is a crucial part for verification of the driver using past trips data.
2) We used only the data of the vehicle status that could be extracted from On-board diagnostics II (OBD) port. Furthermore, the frequency of data collection was 1 Hz, acquiring information at one second per feature. For concrete analysis of differences in the driving patterns of each driver, it is necessary to collect more data within a fixed time. However, we evaluated the driver identification performance using the data collected at 1 Hz.
3) In further experiments and evaluation, we assessed the performance and accuracy considering the time window, training size, number of drivers, types of classification

Byung Il Kwak is with the Graduate School of Information Security, Korea University, Seoul, Republic of Korea (e-mail: kwacka12@korea.ac.kr).

Mee Lan Han is with the Graduate School of Information Security, Korea University, Seoul, Republic of Korea (e-mail: blosst@korea.ac.kr).

Huy Kang Kim is with the Graduate School of Information Security, Korea University, Seoul, Republic of Korea (e-mail: cenda@korea.ac.kr), (Corresponding author) .

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics

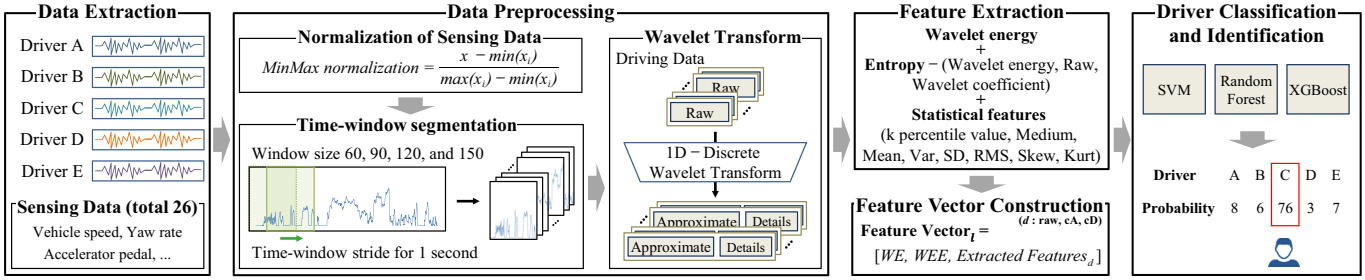


Fig. 1: Overview of driver identification system.

algorithms, and feature importance.

II. Related work

In previous driver identification studies, the type and characteristic features of a dataset, as well as the employed algorithm, are important factors to obtain high identification accuracy. The data type in driver identification studies was divided into three kinds of driving data. First, simulation driving data have been used to overcome the lack of real driving data or to reproduce situations that cannot be demonstrated in an actual driving environment. A simulation program allows one to set the number of vehicles and road conditions [3]. Secondly, the other type of driving data is the output from the OBD-II port in a vehicle, showing the status of the vehicle, such as fuel consumption, throttle position, and other sensing data. These driving data can be extracted from most vehicles equipped with the OBD-II port. These data are also frequently used in studies for the analysis of the driving pattern, as they have better accessibility for interpretation of the vehicle status [4], [5]. Finally, the sensor data refer to the inertial measurement unit (IMU) data. The IMU data are generated from the manifold sensors in mobile devices, i.e., the gyroscope, accelerometer, GPS, and so forth. Owing to the use of the smartphone app, these data are more accessible than different kinds of sensor data [6].

Furthermore, the features of a dataset also play an important role in improving the performance of driver identification. Data output from in-vehicle sensors (e.g., the accelerator and brake pedal) has been used many studies as the features, because they are easily affected by the driver's driving habits. Several research groups determined the need for GPS data to obtain the exact point of the acceleration value generated during turn driving. For driver identification, these methods are particularly effective in a specified section, such as a turn section among various road types [7], [8], [9]. Moreover, driver behavior analysis has been conducted considering the events that occur before driving. These include differences of driver behavior in pre-trip events, such as door opening and closing, starting ignition, and seatbelt fastening. According to earlier research, these differences stem from the differences in the drivers' general habits. Regardless of the road conditions and traffic situation, these features entirely depend on the specific driver's behavior [10].

After completing pre-processing of the data type and feature selection, the classification algorithm for driver identification is typically selected. Among the supervised learning algorithms, the decision tree, Random Forest, and K-Nearest Neighbor (KNN) are typically used to classify drivers. These algorithms are intuitively understandable and straightforward to implement. However, larger amounts of data lead to a slower analysis speed, although a large number of samples yield a higher accuracy. In the case of the Support Vector Machine (SVM) algorithm, it is difficult to employ large-scale training samples, which can lead to both classification and prediction problems. However, due to lower overfitting, SVM has a high accuracy in prediction [11]. In the case of unsupervised learning algorithms, such as the Expectation-Maximization (EM) algorithm, K-means algorithm, and Gaussian Mixture Model (GMM), the algorithms quickly perform clustering for data features without training data after the input of unrefined data. However, these approaches have limitations in terms of grasping the classification criteria from the results and predicting the cluster [12]. Recently, deep learning methods have been actively discussed. For instance, Chen et al. (2019) proposed a driver identification method based on hidden feature extraction. Depending on the construction of a deep non-negativity constrained auto-encoder network, hidden features of driving behavior are extracted automatically to identify the driver [13]. Another study introduced a driver identification model based on the data extracted from smartphone sensors and the OBD-II port. This model derives more stable and reproducible results using the cross-validation technique that employs convolutional neural network (CNN) and recurrent neural networks (long short-term memory) RNN/LSTM [14]. The CNN algorithm handling the Controller Area Network (CAN) bus data yields excellent performance in driver identification [15], [16]. Although it requires large amounts of computation for both the calculation and training of data, it offers highly accurate performance. If the issue of the GPU usage can be overcome, these deep learning algorithms could be applied to a variety of data types.

III. Methodology

In this section, we explain the method of the driver identification system used in the present study. The method comprises four processes to identify the drivers, including (1) data extraction, (2) data preprocessing, (3) feature extraction and feature vector construction, and (4) driver classification and identification (see Fig. 1). First, in the data extraction phase, the status of the vehicle during driving is extracted into the driving dataset,

TABLE I: Vehicle sensing data.

#	Feature name	Related category	Range (unit)	Description
1	Fuel consumption	Fuel	0 – 8388 (mcc)	Fuel efficiency of an engine
2	Accelerator pedal value	Fuel	0 – 99.2 (%)	Degree to which driver is depressing the accelerator pedal
3	Throttle position signal	Fuel	0 – 100 (%)	Value of throttle valve the according to the status of deceleration or acceleration
4	Short term fuel trim bank 1	Fuel	-100 – 99.2 (%)	Immediate correction value determined from the oxygen sensor readings
5	Intake air pressure	Fuel	0 – 255 (kPA)	Pressure of air inhaled to engine
6	Engine soaking time	Fuel	0 – 255 (Min)	Period to get the required amount of fuel for each cylinder
7	Engine fuel cut off	Fuel	0 or 1	Throttle valve closed, and the accelerator pedal released for shutting of injecting the fuel
8	Long term fuel trim bank 1	Fuel	-100 – 99.2 (%)	Correction value being used by the fuel control system in loop modes of operation
9	Vehicle speed	Fuel	0 – 254 (km/h)	A speed value of the vehicle
10	Brake switch ON and OFF	Fuel	0 or 1	A condition of the brake switch
11	Torque of friction	Engine	0 – 100 (%)	A torque caused by the frictional force
12	Current spark timing	Engine	-35.625 – 60 (^{o})	The spark time at the moment of change of the driving status (i.e., acceleration or deceleration)
13	Engine coolant temperature	Engine	-40 – 200 (^{o}C)	Temperature of the engine coolant of an internal combustion engine
14	Engine idle target speed	Engine	0 – 2550 (rpm)	Target speed of the engine idling corresponding to an engine coolant temperature
15	Calculated load value	Engine	0 – 100 (%)	Percentage of peak available torque
16	Minimum indicated engine torque	Engine	0 – 99.61 (%)	Calculated value of minimum torque
17	Flywheel torque	Engine	-3276 – 3276 (Nm)	Power of flywheel that acts as a weighing point for evenly distributing the torque
18	Activation of air compressor	Engine	0 or 1	Value of air compressor's work
19	Calculated road gradient	Engine	-16 – 15.5 (%)	The slope of the vertical alignment of a road
20	Torque convertor speed	Transmission	0 – 16383 (rpm)	A particular kind of fluid coupling used to transfer rotating power from a prime mover
21	Current gear	Transmission	0 – 8	Current step of gear
22	Transmission oil temperature	Transmission	-40 – 215 (^{o}C)	A fluid temperature inside the transmission
23	Clutch operation acknowledge	Transmission	0 or 1	A condition of clutch's operation
24	Steering wheel speed	Wheel	0 – 1016 (^{o}/s)	A speed value of steering wheel
25	Steering wheel angle	Wheel	-3276 – 3276 (^{o})	A value of steering wheel angle
26	Yaw rate	Wheel	-2 – 2 (deg/s)	Vehicle rotation speed

and important data required for classification are selected. The sub-phases, such as time window segmentation and wavelet transform, are performed to obtain the characteristics of the driving pattern. In the feature extraction phase, crucial features of the driving pattern are extracted. Then, all features are used as input values into the classification algorithms, such as the SVM, Random Forest, and XGBoost. Finally, the results of the classification presented as the probability to identify the authorized driver are used to determine whether or not the actual driver is the vehicle owner.

A. Data Extraction

This section describes the process of data extraction from the vehicle status information during driving. The modern vehicle comprises numerous sensors for automatic control without manual adjustment. These sensor data are essential not only for vehicle maintenance, but also for driver identification. In the present study, we selected and extracted 26 sensing datasets in the data extraction phase.

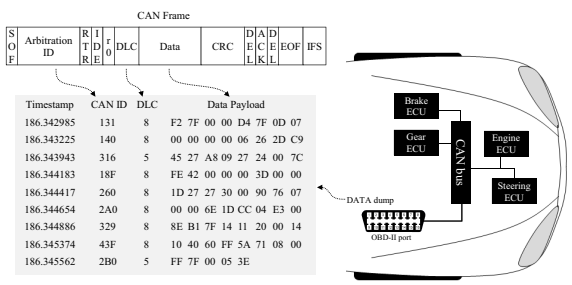


Fig. 2: CAN data extraction.

1) OBD-II Data Extraction: Modern vehicles are equipped with Electronic Control Unit (ECU) that exquisitely controls the internal combustion engine and vehicle body. The embedded ECU and multiple sensors observe the many information in real-time to recognize the surrounding circumstances of driving. The observed sensing data heads to the main ECU through the CAN bus. The observed data in the CAN bus consists of a CAN message format, including CAN id, data length code (DLC), and data payload (see Fig. 2). What we used as the OBD-II data was the response data about the message of the diagnostic request to the target ECU. As self-diagnostic and reporting, OBD-II monitors functional abnormality of the vehicle system [17]. OBD-II data manifest the vehicle status through embedded sensors as a numeric or nominal value. In order to obtain the sensing data, we extracted the driving data using an OBD-II scanner and a laptop computer. The OBD-II scanner was connected to the experiment vehicle, HYUNDAI Sonata 2010. To collect OBD-II data, the laptop computer was linked to the OBD-II scanner via Bluetooth. Data collection was performed during driving, and driving data from all sensors were recorded every second.

2) Selection of Sensing Data: Typically, modern cars have numerous embedded sensors employed to report the status and support vehicle operation, such as a tire pressure monitoring system, steering wheel angle sensor, and brake pedal sensor. In the OBD-II data of the in-vehicle sensors, 'Steering wheel angle' and 'Accelerator pedal value' were commonly used in previous studies of driver identification. The value of the 'Steering wheel angle' shows the driving characteristics through the radius size in turn section of the roads. The other type of sensing data, 'Accelerator pedal value,' shows if the driver has an aggressive driving style. In particular, the value of 'Fuel consumption' changes depending on the

degree of 'Accelerator pedal value.' In this way, the driver's driving characteristics affect, directly and indirectly, fuel-related, engine-related, transmission-related, and wheel-related sensing data. Despite the existence of these useful sensing data, some sensing data could impede driver identification. Furthermore, due to data redundancy and connectivity between the sensors, parts of the sensing data may have a high correlation coefficient. To obtain refined and useful driving data, we eliminated duplicate data from the sensing dataset. In this study, we used the 26 sensing data that do not exhibit redundancy between OBD-II data and could be affected by the driving characteristics (see Table I). Table I provides further detailed description about the sensing data.

B. Data Preprocessing

In this section, we describe the employment of driving data in the classification algorithm. In the data-preprocessing phase, the dataset undergoes three sub-processes: (1) normalization of the sensed data, (2) time-window segmentation, and (3) wavelet transform.

1) Normalization of Sensing Data: Classification algorithms calculate the features' values to get the weights of the model in the training and testing phase. In the process of modeling, the range of non-normalized features generates biased results. Said differently, if each feature (i.e., 'Vehicle speed' and 'Fuel consumption') has a different value range (see Table I), they can be considered as noise in the dataset. To reduce the noise of feature values, we conducted feature normalization of min-max using Eq. (1). The sensing data pass through the data normalization, and the normalized values are scaled from 0 to 1. In Eq. (1), x_i is an *i*th feature (*i* = 1, 2, 3,. . . , n); *min()* and *max()* functions are the minimum value and the maximum value of feature x_i, respectively.

$$min - max\ normalization = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{1}$$

2) Time-Window Segmentation: The driving data generated by various conditional circumstances, such as traffic status, weather, and driving style, can differ in terms of driving time and vehicle status even if the driver and route are identical. For the analysis of driving data with an equal criterion, the analysis of driving patterns should be executed at equal time windows, i.e., within a fixed period. Accordingly, we divided the driving dataset into bits of the uniform size, namely 60, 90, 120, and 150 seconds. Furthermore, we updated the unit of the time window as a one second, which allows for the detailed analysis of changes in the sensing data values.

3) Wavelet Transform: The analysis method should be considered by the type of data to extract the significant driving pattern. The driving data we collected from the vehicle could be considered to be time-series data. Overall, two types of transform—Fourier transform and wavelet transform—are used as a signal analysis tool for frequency analysis. In time-series analysis, even if Fourier transform could be appropriate for frequency analysis, its limitation is that it does not consider the loss for time information. However, the wavelet transform is appropriate to cover the time and frequency information [18]. With consideration of time-variant, a solution to analyze the time-series data can be a wavelet transform. Furthermore, the wavelet transform is particularly useful for non-stationary data, such as time-series data [19]. For that reason, we applied the wavelet transform that can consider not only time, but also frequency.

In general, wavelet transform is a powerful tool for signal analysis and image processing that works through shifting and scaling of the signal. In wavelet analysis, the Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT) are mainly used in various fields of research, including noise reduction, image compression, and signal processing. The important difference between CWT and DWT is the amount of calculation for scale and location factors. CWT calculates in as many scale and location factors as possible, whereas the DWT calculates only some coefficients, excluding redundant scales and coefficients [20]. This calculation of redundancy means that CWT could require more computational resources than DWT, whereas DWT could have a loss of information as compared to CWT. However, DWT provides sufficient information for an in-depth analysis of the original signal. Accordingly, we used DWT that can not only reduce the time complexity and the computing resources, but also provide sufficient information for the data analysis without the duplication of location factors. DWT is defined as follows (see Eq. (2)):

$$DWT(j,k) = \frac{1}{\sqrt{|2^j|}} \int_{-\infty}^{\infty} x(t)\psi(\frac{t - k2^j}{2^j})dt \tag{2}$$

where 2^j and $k2^j$ are the scaling and time-shifting parameters on powers of two for the DWT, respectively.

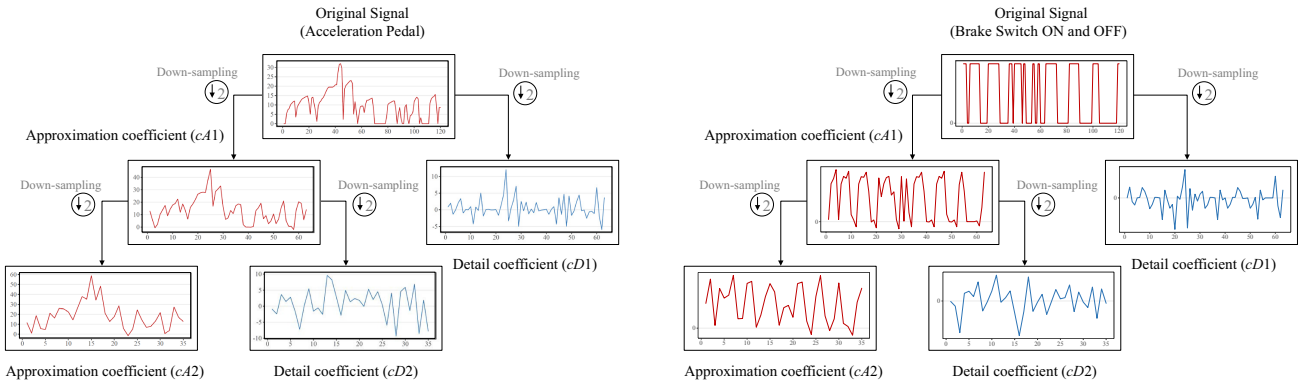
$$y_{LPF}(k) = \sum\nolimits_n x(t) \cdot h_n(2k - n) \tag{3}$$

$$y_{HPF}(k) = \sum\nolimits_n x(t) \cdot g_n(2k - n) \tag{4}$$

The original signal is *x(t)*, and *n* is an integer. Signal *x(t)* can be decomposed by filters, namely a High-Pass Filter (HPF) and a Low-Pass Filter (LPF) in the DWT process. h_n and g_n depict LPF and HPF, respectively. If signal *x(t)* passes through the HPF and LPF, the original signal is decomposed into the approximation coefficients (cA1, $y_{LPF}(k)$) and detail coefficients (cD1, $y_{HPF}(k)$) as the output of the two filters in (3) and (4). We employed the db4 (Daubechies 4) wavelet in various filters.

Furthermore, for a more detailed analysis, the wavelet transform was applied in conjunction with multi-resolution decomposition. To show the wavelet transform of the sensing data, we displayed the multi-resolution decomposition of 'Accelerator pedal value' in Fig. 3a. The original sensing data were decomposed into level 1 approximation coefficients and level 1 detail coefficients through the wavelet transform. Similarly, the level 1 approximation coefficients were decomposed into level 2 approximation coefficients and level 2 detail coefficients. These decomposition processes were repeated until the maximum level was achieved. In our experiments, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics



(a) Multi-resolution of 'Acceleration Pedal Value'

(b) Multi-resolution of 'Brake Switch ON and OFF'

Fig. 3: Multi-resolution decomposition tree to level 2 depending on the data type.

maximum level of the decomposition was set to two, three, and four according to the length of the input signal 60, 90, 120, and 150 seconds, respectively. As shown in Fig. 3b, we depicted the multi-resolution decomposition of 'Brake Switch ON and OFF' with a binary value type from original sensing data to level 2 approximation and detail coefficients in the same way for 'Accelerator pedal value.' To improve the accuracy of driver identification, we applied the results of the wavelet transform, such as approximation and detail coefficients, along with the original sensing data.

C. Feature Extraction and Feature Vector Construction

In the present study, we extracted three types of features—namely, the entropy value, wavelet energy, and statistical features—in order to obtain informative values from the sensing data. Before feature extraction, we set the time window to account for the time-series data, including approximation coefficients, detail coefficients, and original sensing data called as raw data. The time-series data are needed a preprocess to extract and to analyze the driving pattern. The wavelet features and statistical features show the representation of the time-series as a specific value. The value of representation data presents a summary value of the time-series data, which shows driving characteristics as a specific value; the summary values allow for an easy analysis of the driving pattern. Accordingly, we extracted the wavelet energy, the summary value of the entropy, and statistical value in time-series data (see Table II).

1) Wavelet Energy and Wavelet Energy Entropy: We considered the wavelet energy (WE) and wavelet energy entropy (WEE) to analyze the variations in the driving patterns. In order to extract WE and WEE, we obtained the wavelet coefficients—namely, the approximation coefficient and detail coefficient, i.e., cD1 – cD4 and cA1 – cA4, through the multi-resolution decomposition. As shown in Table II, the C_j was the coefficients for the jth resolution level, and k was each sampled time, the wavelet energy set in the all resolutions level is E_{tot}. The WEE was obtained by the entropy function of the relative wavelet energy [21], [22]. The entropy function of wavelet energy is a Shannon entropy that provides an expected value of the information set. The WE and the WEE value can supply the information on the driver's driving pattern, which is variation or consistency. For instance, the aggressive driving style can be represented by the WE of 'Accelerator pedal value.' The wavelet energy refers to the degree of the signal variation, and the result, i.e. the WE value of 'Accelerator pedal value', shows the aggressive driving style in terms of vehicle acceleration.

2) Shannon's Entropy and Statistical Feature: To obtain a summary of driving patterns, we considered the statistical features and the entropy value. Statistical features represent the summary of the information within the fixed time window, which is an efficient method to learn about the differences between drivers' driving patterns. For example, root mean square (RMS) and variance on 'Fuel consumption' show whether the driver's driving style is eco-drive or not. Furthermore, the statistical features require fewer computational resources than the use of all features. The entropy represents the attribute for the quantities of information obtained from the driving data through a random variable. To include entropy into the

TABLE II: Feature of extraction from original vehicle feature.

#	Feature name	Description
1	Wavelet energy	$E_j = \sum_k \lvert C_j(k)\rvert^2,\ C_j(k) =< X, \varphi_{j,k} >$
2	Wavelet energy entropy	$WEE = -\sum p_j \cdot ln[p_j],\ p_j = \frac{E_j}{E_{tot}}$
3	Data entropy	$H = -\sum_{i=1}^{n} p_i \cdot ln[p_i],$ $where\ p_i = \frac{y_m(i)}{\sum_{i=1}^{n} y_m(i)}$
4	The k percentile value (k is 5, 25, 75, 95)	$P_k = \frac{k(N+1)}{100}$
5	Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x(i)$
6	Median	$Median = \frac{n+1}{2}$
7	Variance	$Var = \frac{1}{N}\sum_{i=1}^{N} (x(i) - \bar{x})^2$
8	Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x(i) - \bar{x})^2}$
9	Root mean square	$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x(i)^2}$
10	Skewness	$Skew = \frac{1}{N}\sum_{i=1}^{N} (\frac{x(i) - \bar{x}}{\sigma})^3$
11	Kurtosis	$Kurt = \frac{1}{N}\sum_{i=1}^{N} (\frac{x(i) - \bar{x}}{\sigma})^4$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics

feature vector, we measured the information probability of the distribution for the segmented dataset. The features we used in our method are summarized in Table II. The statistical features included a total of 11 features, such as the k percentile (where k is 5, 25, 75, 95), mean, median, variance, standard deviation, RMS, skewness, and kurtosis.

The construction process of the feature vector was as follows:

Step 1. All 26 features were subjected to the min-max normalization to reduce classification errors.

Step 2. Normalized feature data were allocated into bits of equal length depending on the size of the window (60, 90, 120, and 150 seconds).

Step 3. Allocated data of the same length were decomposed into the approximation coefficients and the detail coefficients. Raw data before the wavelet transform, approximation coefficients, and detail coefficients data were applied in the next phase to extract summary values.

Step 4. The extracted features (i.e., Shannon's Entropy and Statistical Feature) were obtained for all three data chunks. As an exception, the WEE values were extracted only once from the raw data, which is resulted from the set of wavelet energy. The WE was acquired per one by decomposition level and each coefficient. Moreover, the WE value was measured until the maximum level of the decomposition according to the length of raw data was reached.

Step 5. When l was vehicle sensing data in Table II, the extracted features were used to construct the feature vector using Eq. (5) and (6) given below:

$$Extraced\ features = (Shannon's\ Entropy,\ The\ k\ Percentile\ (k\ is\ 5,\ 25,\ 75,\ 95)),\ Mean,\ Median,\ Variance,\ SD,\ RMS,\ Skewness,\ Kurtosis) \quad (5)$$

$$Feature\ vector_l = [WE,\ WEE,\ Extracted\ features_{raw},\ Extracted\ features_{cA},\ Extracted\ features_{cD}] \quad (6)$$

D. Driver Classification and Identification

In this phase, we employed three classification algorithms—namely, SVM, Random Forest, and XGBoost. In what follows, we describe the driver identification process.

1) Driver Classification: Three algorithms were used to resolve the problem of classification using supervised learning, and each algorithm has the characteristics necessary to solve the problem. SVM is a model employed for classification and regression analysis. The algorithm determines a hyperplane to discriminate each different set and to have the farthest distances from the data of each class. SVM can select the kernel function depending on the problem. Furthermore, the algorithm has an excellent classification performance [23]. Random Forest is one of the ensemble learning algorithms and a combination of many decision trees. Furthermore, the general property of Random Forest is a voting algorithm in terms of numerous generated decision trees. In this way, the algorithm has the advantages of high accuracy and mitigation of the over-fitting problem [24]. The XGBoost is another ensemble learning algorithm employing boosting. The boosting consists of various models and applies the misclassified data in the current model as input to the next model with weights. Due to the characteristics of parallel computing, regulation, and tree pruning process, XGBoost exhibits excellent performance in solving the problem of binary and multi-classification [25].

To achieve a high accuracy, we selected the aforementioned three algorithms known for their excellent classification performance. After construction of the feature vector, it was considered as the input value in the driver classification section. After classification, we obtained the probabilities of a particular driver driving the vehicle, which depended on the algorithm.

2) Driver Identification: In the driver identification step, we decided whether or not the driver was the vehicle owner based on the so-called driver probabilities. Driver probabilities vary considerably depending on the criteria of the selection and classification algorithms. For example, if the identification probability is 99%, the driver is almost definitely the vehicle owner. In the case of probability of 60%, the driver identification could be more difficult. The differences in the results from the three classification algorithms impose limitations on specifying the threshold value that distinguishes drivers in these systems. To address these concerns, we defined criteria upon which a driver with a high accuracy of the classification would be the authenticated driver. For instance, if the accuracy of the classification for driver A was 90%, driver B was 7%, and driver C was 3%, then the result was determined to be Driver A in the driver identification step.

IV. Evaluation

In this section, we discuss the experimental setup, performance of the driver classification and identification, effects of rate in training and testing size, driver identification in short-term data, driver identification in urban road, feature importance of driver identification, and driver characteristic perception based on the results of feature analysis.

A. Experimental Setup

To evaluate the accuracy of driver identification, we collected various sensing values from the vehicle's built-in

Fig. 4: The motorway of driving route in experiment (distance: about 17 km).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics

TABLE III: Participants information

Driver	Gender	Age	Driving career
A	Male	26	3 years over
B	Male	30	1 year over
C	Male	26	3 years over
D	Male	27	3 years over
E	Female	39	1 year over

sensors during the participants' driving. OBD-II data were captured every second to track the driver during driving. Many sensors were employed to measure the driving data; however, to grasp the effects of in-vehicle sensors, we composed the features as the only in-vehicle sensor without other external sensors, such as wearable devices, GPS trackers, and smartphones. Five participants (four men and one woman) participated in our experiments. Table III shows the participants' information that can make the distinct differences in driving pattern of a driver. The participants are categorized depending on gender, age, driving career. Four males and one female participant drove the car, and the driving career was between one year and three years. The participants drove the HYUNDAI Sonata 2010. Fig. 4 shows the experimental route (the length was about 17 km). The course consisted of two types of roads: a typical urban road and a motorway. In this experiment, all participants drove the same route ten times. In order to control the variable about the driving, the driving conditions were set up the conditions, including driver's physical condition (low-risk of the driving), similar weather (sunny day), similar time slot of the driving (10-12 AM, 2-5 PM). Then, we divided the driving data into nine trips for the training dataset and used one trip as the test data; we also selected the test data as one driving data for all participants. The driving dataset for the training and test is shown in Table IV. To apply the algorithms (i.e., Random Forest, SVM, and XGBoost), we used the WEKA[1] and XGBoost Python library as a platform to apply the algorithms. Furthermore, we set the hyper parameters as a default about the three algorithms: batch size '100', complexity parameter '1.0', kernel 'polynomial kernel' in SVM. Similarly, the batch size '100', depth 'unlimited,' and number of iterations '100' were set as in Random Forest. The batch size '1000', maximum depth of a tree '6', and learning rate '0.1' were set in XGBoost.

B. Performance of Driver Identification

The purpose of driver identification is to accurately determine whether or not the driver driving the vehicle is an authorized vehicle driver/owner. Data from past driving trips summarizes current driving trips for a specific driver. All driving data were segmented to a fixed size of the time window and then labeled with a class name. The accuracy was chosen to evaluate the performance of our proposed model. The driver identification is a problem of multi-class classification. Therefore, we considered a true positive (TP) that indicated that the driver was correctly classified. The aforementioned performance metric was obtained using Eq. (7).

$$Accuracy_c^t = \frac{True\ Positive_c^t}{Number\ of\ segmented\ data_c^t}, \quad (where\ c\ is\ class,\ i.e.,\ A,\ B\ ,C\ ,D,\ and\ E) \tag{7}$$

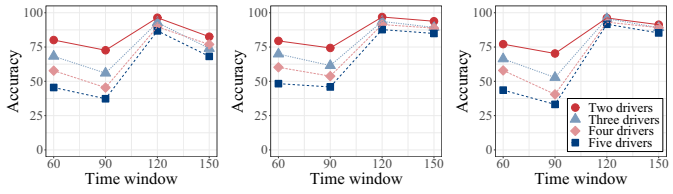
Here, t is the size of the time window, and c is the class. The accuracy of the driver classification is shown in Fig. 5. The accuracy of driver classification differed for each of the classification algorithms—i.e., SVM, Random Forest, and XGBoost. The accuracy was the highest in the time window of 120 s, and the XGBoost algorithm showed the highest accuracy (91.6%) in the classification for five drivers. Furthermore, to evaluate the effects of the time window, we assessed the accuracy of driver identification with respect to the time window length. Hence, we tested the driver identification by adjusting the length of the time window to 60, 90, 120, and 150 seconds. With the growth of the time window, the 120 s length of the time window was found to be the most adequate (See Fig. 5). The accuracy of the time windows 60 s and 90 s was lower than the accuracy of the time window at 150 seconds. Regarding the number of drivers, the result suggested that fewer drivers yield better accuracy.

To clearly identify the driver, classification results were provided as probabilities that the driver was the correct person. Fig. 6 presents the driver identification results for the three algorithms. Fig. 5a shows that all drivers were identified with certainty. In the SVM algorithm, drivers C, D, and E were

[1]The opensource software for machine learning

TABLE IV: Driving dataset of the training and the test

Driver	Training dataset		Test dataset	
	# of records	# of trip	# of records	# of trip
A	22,853	9	2,209	1
B	19,833	9	3,105	1
C	18,105	9	2,234	1
D	19,477	9	2,481	1
E	22,133	9	2,597	1
Total	102,401	45	12,626	5



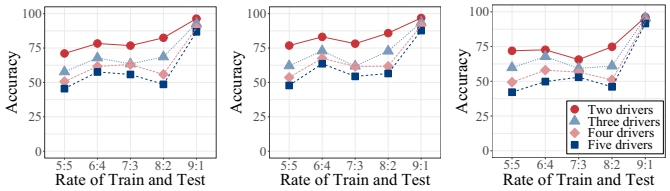
(a) SVM (b) Random Forest (c) XGBoost

Fig. 5: Accuracy of driver classification for each algorithm with respect to time window length.

(a) SVM (b) Random Forest (c) XGBoost

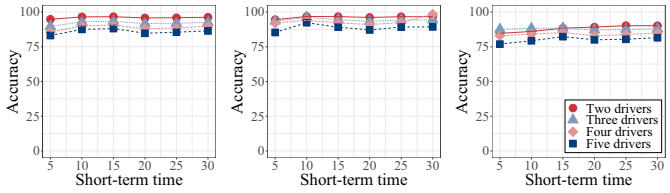
Fig. 6: Confusion matrix of driver identification.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics



(a) SVM (b) Random Forest (c) XGBoost

Fig. 7: Accuracy of driver classification for train and test dataset size.



(a) SVM (b) Random Forest (c) XGBoost

Fig. 8: Accuracy of driver classification in short-term data.

identified with the accuracy of 94% or higher, while drivers A and B had the accuracy of 72%. As shown in Fig. 5b, drivers B, C, and E were identified with the accuracy of over 92%, while driver A had identification accuracy of about 61%, and driver D had about 81%. Similarly, drivers B, C, D, and E were identified with the resulting accuracy of over 97% in XGBoost, while the accuracy for driver A was only 57% (Fig. 5c). However, as driver identification, the probabilities of the result from the three classification algorithms present that the driver identification was performed correctly.

C. Effect of Rate in Training and Testing Size

In driver identification, the size of the training dataset is crucial, as the size of the learning data could have limitations in real conditions. If the size of the training data is too small, the performance of the driver identification could be undesirable in the test phase. By contrast, if the size of the training data is too large, the performance is better; however, enormous training and computer resources are needed to handle it. To obtain the relevant ratio of dataset about the training and testing, we adjusted the ratio of the training and test data to 9: 1, 8: 2, 7: 3, 6: 4, and 5: 5 based on the number of driving trips. For example, the ratio of 9: 1 indicates that the training data included nine trips, while the testing data included only one trip. Fig. 7 presents the accuracy of driver classification regarding various sizes of the training data in the three algorithms. The suitable ratio of the training and test data was 9: 1 for driver identification, while the 5: 5 ratio showed the lowest accuracy in the three algorithms. In some ratios of the train and test data, the accuracy fluctuated among the three algorithms. Furthermore, the accuracies of driver identification for three and four drivers fluctuated regardless of any trends in the training and test data ratio 8: 2, 7: 3, and 6: 4. The results showed the number of minimum training datasets to obtain acceptable results.

D. Driver Identification in Short-term Data

To validate short-term data, we made an experiment of the driver identification that had short-term test data. The data consisted of the short-term test data (from 5 min to 30 min by 5 min unit). In this experiment, we applied the 120-sec size of time-window, which has the best performance in Fig. 5. The classification results that are 2-4 driver classification are shown in Fig. 8, the X-axis is short-term data (from 5 min to 30 min), and the Y-axis is accuracy. In Fig. 8, the identification performance of two drivers for short-term 5 min shows almost 95% in SVM and Random Forest, and the performance of XGBoost was almost 87%. In addition, the accuracies of five drivers for the same short-term were lower than the accuracy of two drivers' identification. In these results, there is a small difference in performance as compared with the performance of full test data about SVM and Random Forest. In XGBoost, the results have significant differences between 10-15% short-term 5 min and full test data. However, the performance of XGBoost gradually increased by getting more test data; the accuracy in the full test was the highest. From the results of the short-term test, it can be concluded that the proposed method could be appropriate for driver identification in short-term data.

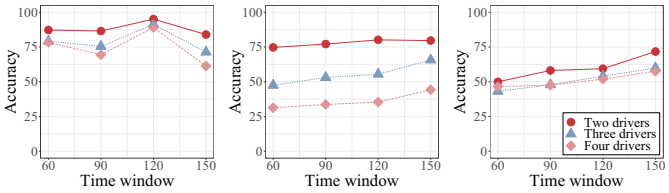
E. Driver Identification in Urban Road

In this section, we evaluate the new driving data from the driving section of the urban road. In this section, to learn the applicability of other driving sections, we experimented with the driver classification of the driving data from the section in the urban road. We extracted the driving data for four drivers in the driving section—namely, urban road 1 and urban road 2 (see Fig. 9). The four participants joined in this experiment with the vehicle (HYUNDAI Sonata 2010). Each participant drove the urban road 1 and urban road 2 from ten to twelve times. We used one trip as the test data and the rest trip data as the training data. Similarly, the length of time-window was adjusted to 60, 90, 120, and 150 seconds.

As can be seen in Fig. 10, the classification results show the best in the SVM algorithm. Furthermore, the results showed

Fig. 9: The urban road of driving route in experiment; Urban road 1 (blue line of left side) distance: 5.5km, Urban road 2 (red line of right side) distance: 8.4km).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics



(a) SVM (b) Random Forest (c) XGBoost

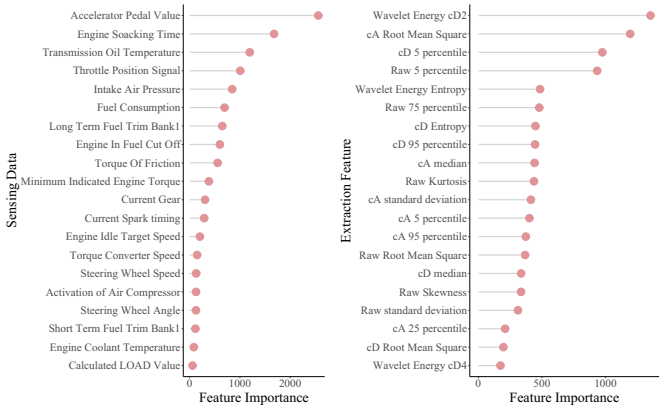
Fig. 10: Accuracy of driver classification in urban road.

that the more number of drivers, the lower the accuracy of driver classification. Differently from the results about the motorway, the accuracy of driver classification was relatively low in Random Forest and XGBoost. In the time-window of 120 s, the accuracy of SVM with the two drivers was 95.07%, 91.36% with the three drivers, and 89.06% four drivers. In this experiment, the accuracies of Random Forest and XGBoost increased by the raising of the time-window size. According to these results, in order to obtain a high accuracy, Random Forest and XGBoost algorithms could be appropriate in a sufficient length of the time-window, which could show in the differences of driver's driving pattern. Moreover, the merged driving data would not be appropriate to classify the driver in the urban road. Furthermore, since the urban road has various obstacles, including intersections, traffic lights, limitations of vehicle speed, and the short driving section between crossroads, it would be difficult to figure out the characteristic of the driving pattern in the urban road.

F. Feature Importance of Driver Identification

The feature importance indicates the level of discrimination in driving behavior, and it helps choose crucial and useful features in a pool of various features. To be specific regarding the impact of the features, we performed an experiment of the feature importance in the two views of the features.

1) Feature Importance based on Sensing Data: Various sensors in the vehicle generate many types of sensing data while the vehicle is in the driving mode. With diverse data, the selection among sensing data is essential to determine the



(a) Sensing Data (b) Extracted Features

Fig. 11: Feature importance.

TABLE V: Rank of feature combination in the sensing data and the extracted features; A feature in parentheses indicates a sensing data

Rank	Extracted Features
1	cD-5 percentile (Accelerator pedal value)
2	cD2-Wavelet energy (Accelerator pedal value)
3	cA-Root mean square (Accelerator pedal value)
4	Raw-5 percentile (Intake air pressure)
5	cD2-Wavelet energy (Engine soacking time)
6	cA-Root mean square (Fuel consumption)
7	Raw-Kurtosis (Engine soacking time)
8	cD-Median (Engine soacking time)
9	cA-Median (Accelerator pedal value)
10	Raw-Root mean square (Transmission oil temperature)

features that influence driver identification. Fig. 11a shows the feature importance in the top 20 of sensing data. The 'Accelerator Pedal Value' was found to be the best sensing data feature to discriminate the driving pattern, while the 'Engine Soaking Time' had a lower impact. Direct effects of the driving (i.e., handling the steering wheel, pushing the accelerator pedal) could be significant features for the determination of the driving pattern depending on the driver; however, the results differed from than the expectation (see Fig. 11a). For instance, the 'Steering wheel angle' and 'Steering wheel speed' had a low feature importance for our experiment. This result could be related to the differences between the urban road and the motorway. First, unlike the motorway, the urban road has many crossroads and turns. Furthermore, the urban road has traffic lights that are not typically present in the motorway. Finally, for the safety of pedestrians, the urban road has a lower upper speed limit than the motorway. These differences can influence the significance of different features for driver identification.

2) Feature Importance based on Extracted Features: To determine the impact of the extracted feature, we measured the feature importance top 20 of extracted features (Fig. 11b). The features related to wavelet coefficients were found to have a high feature importance, meaning that the wavelet transform played an important role in driver identification. For example, the 'Wavelet Energy cD2' is the wavelet energy in detail coefficients of the two-level decomposition. The size of the change from the sensing feature, like the dynamic change of the accelerator pedal sensor, is the best significant feature to classify the driving patterns.

The ranking of the feature combination of sensing data and extracted features is shown in Table V. The best combination was 'cD5 percentile' and 'Accelerator pedal value,' followed by 'cD2 Wavelet energy' and 'Accelerator pedal value.' In extracted features, the features through the wavelet transform, such as detail coefficients or approximation coefficients, had a higher feature importance value than raw data. Moreover, the 'Accelerator pedal value' each driver showed the best characteristics that could be differentiated each of them in sensing data. The results of the analysis of feature importance highlight the importance of features for driver identification. In this section, we determined the significance of sensing data (i.e., 'Accelerator pedal value,' 'Engine soaking time,' 'Transmission oil temperature,' and so forth.) via the analysis of feature importance. Based on these results, we learned about

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.2999911, IEEE Transactions on Industrial Informatics

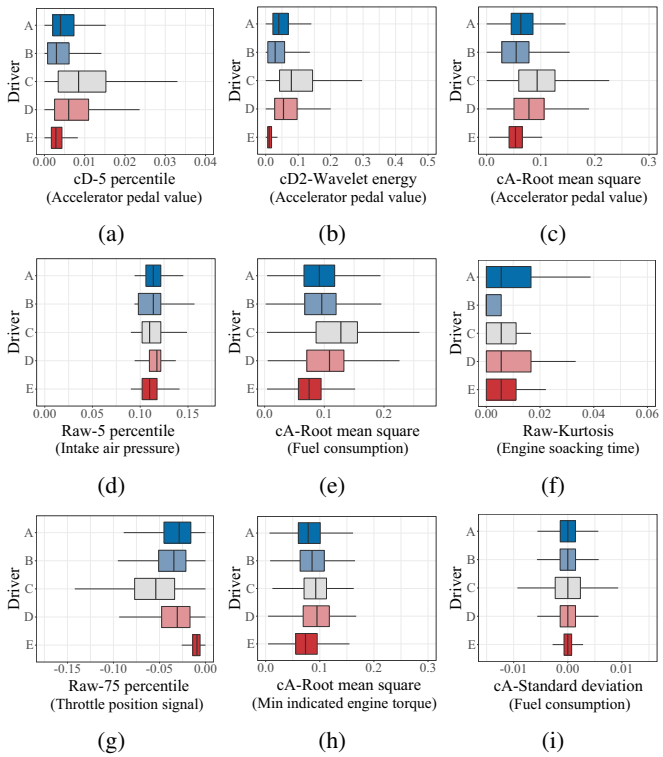


Fig. 12: Characteristics of driving patterns based on extracted features; A feature in parentheses indicates a sensing data.

the importance of the wavelet transform for the performance increase in driver identification. Moreover, we managed to determine the feature combination in the sensing data and extracted features.

G. Driver Characteristic Perception based on the Feature Analysis

To evaluate the characteristics of the driving pattern, we analyzed the extracted feature using the boxplot presentation (Fig. 12). The X-axis in Fig. 12 depicts the extracted feature value, while the Y-axis is a label of the true driver. Fig. 12 shows that driver C and E had significantly different driving patterns in most sub-figures of Fig. 12. Fig. 12a depicts a boxplot for the 'cD-5 percentile (Accelerator pedal value),' revealing five percentiles for the detail coefficients of the driver's 'Accelerator pedal value.' Drivers C and E exhibited a significant difference in Fig. 12a. In this way, we could determine the driving style of driver C, who had an irregular driving habit with regards to accelerating the vehicle. In contrast, driver E had a narrower distribution of accelerations than driver C (see Fig. 12a and 12b). Fig. 12b shows the box plot of the 'cD2-Wavelet energy (Accelerator pedal value)', where driver C showed a more stable driving style than driver D. In Fig. 12h and 12i, driver C had a broader range than driver E with regard to the 'Throttle position sensor' and 'Fuel consumption.' Consequently, the driving style of driver C was more aggressive and uneven than that of driver E. Based on these differences in driving patterns, the driver can be reliably identified.

V. Conclusion

In the present study, we proposed a method of driver identification based on the wavelet transform for driving-behavior analysis. To this end, we first conducted data pre-processing including feature normalization, feature extraction, and feature vector construction. The aim of the data pre-processing was to reduce noise and eliminate incomplete data. Then, we performed and evaluated driver identification using three different machine-learning algorithms based on the wavelet transform. This process was performed by changing the length of the time window of data for classification analysis. The results obtained by applying the proposed method demonstrate a high accuracy of driver identification in the trip-based driving data. Previous studies had the limitation of applying random sampling on the same trip. However, our evaluation results demonstrate that the proposed method based on the wavelet transform can overcome the limitations of driver identification for the new trip. Moreover, our research makes a new contribution by suggesting a structured process to build the driver identification model, including the wavelet transform, the machine learning algorithms, and other processes. We applied the data normalization, time-window segmentation, feature extraction, and feature vector construction. The roles of processes included the reduction of bias and noise, the data segmentation about the proper size, and the significant feature extraction to obtain the distinctive driving pattern.

Driver identification is highly applicable in various aspects of the vehicle industry, including customized driver personalization services and driver authentication. In future research, it would be necessary to extend our findings by performing a cross-analysis with multiple behavioral patterns (i.e., interaction of IoT devices in the vehicle, bio-signal during driving, etc.).

Acknowledgment

This work was supported by Samsung Research Funding & Incubation Center for Future Technology under Project Number SRFC-TB1403-51.

References

[1] C. Miller and C. Valasek, "Adventures in automotive networks and control units," *Def Con*, vol. 21, pp. 260–264, 2013.

[2] C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," *Black Hat USA*, vol. 2015, p. 91, 2015.

[3] Z. Halim, R. Kalsoom, and A. R. Baig, "Profiling drivers based on driver dependent vehicle driving features," *Applied Intelligence*, vol. 44, no. 3, pp. 645–664, 2016.

[4] B. I. Kwak, J. Woo, and H. K. Kim, "Know your master: Driver profiling-based anti-theft method," in *2016 14th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 2016, pp. 211–218.

[5] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, pp. 34–50, 2016.

[6] C.-H. Yang, D. Liang, and C.-C. Chang, "A novel driver identification method using wearables," in *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2016, pp. 1–5.

[7] I. D. Markwood and Y. Liu, "Vehicle self-surveillance: Sensor-enabled automatic driver recognition," in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM, 2016, pp. 425–436.

[8] D. Chen, K.-T. Cho, and K. G. Shin, "Mobile imus reveal driver's identity from vehicle turns," *arXiv preprint arXiv:1710.04578*, 2017.

[9] D. Hallac, A. Sharang, R. Stahlmann, A. Lamprecht, M. Huber, M. Roehder, J. Leskovec *et al.*, "Driver identification using automobile sensor data from a single turn," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016, pp. 953–958.

[10] G. Kar, S. Jain, M. Gruteser, J. Chen, F. Bai, and R. Govindan, "Predriveid: pre-trip driver identification from in-vehicle data," in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*. ACM, 2017, p. 2.

[11] A. Burton, T. Parikh, S. Mascarenhas, J. Zhang, J. Voris, N. S. Artan, and W. Li, "Driver identification and authentication with active behavior modeling," in *2016 12th International Conference on Network and Service Management (CNSM)*. IEEE, 2016, pp. 388–393.

[12] A. Wahab, C. Quek, C. K. Tan, and K. Takeda, "Driving profile modeling and recognition based on soft computing approach," *IEEE transactions on neural networks*, vol. 20, no. 4, pp. 563–582, 2009.

[13] J. Chen, Z. Wu, and J. Zhang, "Driver identification based on hidden feature extraction by using adaptive nonnegativity-constrained autoencoder," *Applied Soft Computing*, vol. 74, pp. 1–9, 2019.

[14] A. El Mekki, A. Bouhoute, and I. Berrada, "Improving driver identification for the next-generation of in-vehicle software systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7406–7415, 2019.

[15] D. Jeong, M. Kim, K. Kim, T. Kim, J. Jin, C. Lee, and S. Lim, "Real-time driver identification using vehicular big data and deep learning," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 123–130.

[16] Y. Xun, J. Liu, N. Kato, Y. Fang, and Y. Zhang, "Automobile driver fingerprinting: A new machine learning based authentication scheme," *IEEE Transactions on Industrial Informatics*, 2019.

[17] Wikipedia, "Obd-ii pids," https://en.wikipedia.org/wiki/OBD-II_PIDs, available: 2020-01-15.

[18] S. Pittner and S. V. Kamarthi, "Feature extraction from wavelet coefficients for pattern recognition tasks," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 1, pp. 83–88, 1999.

[19] L. Chun-Lin, "A tutorial of the wavelet transform," *NTUEE, Taiwan*, 2010.

[20] A. Dainotti, A. Pescapé, and G. Ventre, "A cascade architecture for dos attacks detection based on the wavelet transform," *Journal of Computer Security*, vol. 17, no. 6, pp. 945–968, 2009.

[21] O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, and E. Başar, "Wavelet entropy: a new tool for analysis of short duration brain electrical signals," *Journal of neuroscience methods*, vol. 105, no. 1, pp. 65–75, 2001.

[22] C. Chen and J. Zhang, "Wavelet energy entropy as a new feature extractor for face recognition," in *Fourth International Conference on Image and Graphics (ICIG 2007)*. IEEE, 2007, pp. 616–619.

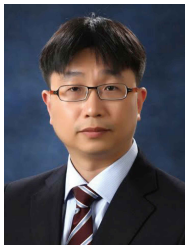
[23] Z. Tian, C. Luo, J. Qiu, X. Du, and M. Guizani, "A distributed deep learning system for web attack detection on edge devices," *IEEE Transactions on Industrial Informatics*, 2019.

[24] S. A. Shevchik, F. Saeidi, B. Meylan, and K. Wasmer, "Prediction of failure in lubricated surfaces using acoustic time–frequency features and random forest algorithm," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1541–1553, 2016.

[25] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, 2018.

Byung Il Kwak [S'20] received his B.S. degree in Computer and Science from Sejong University in 2012. He is in Ph.D. candidate in School of Cybersecurity, Korea University. His research interests include vehicle security, network security and machine learning.

Mee Lan Han [S'20] received her B.S. degree in Computer and Science from DongDuk Women's University, and an M.S. degree in School of Cybersecurity from Korea University, South Korea, in 2002 and 2014, respectively. She is currently working toward a Ph.D. degree in School of Cybersecurity, Korea University, South Korea. Before joining Korea University, she was a manager of the development part against Chinese-speaking countries in Nexon (2004-2012). Her research interests include vehicle security, network security, cyber threat intelligence, and Data mining.



Huy Kang Kim received his Ph.D. in industrial and systems engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2009. He received an M.S. degree in industrial engineering from KAIST in 2000. He received a B.S. degree in industrial management from KAIST in 1998. He founded A3 Security Consulting, the first information security consulting company in South Korea in 1999. Also, he was a member and the last leader of KUS (KAIST UNIX Society), the legendary hacking group in South Korea. Currently he is a professor in School of Cybersecurity, Korea University. His recent research is focused on solving many security problems in online games based on the user behavior analysis. Before joining Korea University, he was a technical director (TD) and a head of information security department of NCSOFT (2004-2010), one of the most famous MMORPG companies in the world.